US EPA TOXCAST DATA RELEASE DECEMBER 2014 – Summary Files

This file describes the contents of the December 2014 ToxCast data release. The zip file contains the following summary-level files:

```
   [1] "AllResults_cyto_dist_141121.csv"
   [2] "AllResults_fitc_Matrix_141121.csv"
   [3] "AllResults_flags_141121.csv"
   [4] "AllResults_hitc_Matrix_141121.csv"
   [5] "AllResults_l4id_Matrix_141121.csv"
   [6] "AllResults_logc_max_Matrix_141121.csv"
   [7] "AllResults_logc_min_Matrix_141121.csv"
   [8] "AllResults_max_mean_Matrix_141121.csv"
   [9] "AllResults_max_med_Matrix_141121.csv"
  [10] "AllResults_modl_ac10_Matrix_141121.csv"
  [11] "AllResults_modl_acb_Matrix_141121.csv"
  [12] "AllResults_modl_acc_Matrix_141121.csv"
  [13] "AllResults_modl_ga_Matrix_141121.csv"
  [14] "AllResults_modl_gw_Matrix_141121.csv"
  [15] "AllResults_modl_la_Matrix_141121.csv"
  [16] "AllResults_modl_lw_Matrix_141121.csv"
  [17] "AllResults_modl_Matrix_141121.csv"
  [18] "AllResults_modl_rmse_Matrix_141121.csv"
  [19] "AllResults_modl_tp_Matrix_141121.csv"
  [20] "AllResults_spid_Matrix_141121.csv"
  [21] "AllResults_tested_Matrix_141121.csv"
  [22] "AllResults_zscore_Matrix_141121.csv"
  [23] "Assay_Summary_141121.csv"
  [24] "Chemical_Summary_141121.csv"
```

In addition to the above listed files, the ToxCast program also released a MySQL dump file containing all data and a beta version of the R package (tcpl) that interacts with the MySQL database used to process all of the data for this release. For information/data not included in the listed summary files, users will need to download and interact with the MySQL database. We also encourage the database users to utilize the 'tcpl' R package containing numerous queries and functionality for easily loading and visualizing the data. At the bottom of this file is an R script to produce all of the listed files, utilizing the MySQL database and 'tcpl' R package.

All information in the summary-level files is reported at the chemical level. When more than one sample existed for a given chemical-assay pair, logic incorporating the distribution of activity calls, the shape of the curves, the cautionary flags, and the potency across samples was used to select a single sample. For more information, see the 'tcplSubsetChid' function in the 'tcpl' R package.

Each of the matrix files(indicated by "_Matrix" in the name) contain 1860 distinct chemicals by 821 assay endpoints, where each cell contains data for a single chemical-endpoint pair. The first column in the matrix files gives the chemical code; column names correspond to assay endpoint name. The zip files contains matrices for 20 of the variables captured at level 4 and level 5 of the analysis:

[1] "fitc" – the fit category
[2] "hitc" – the activity or hit call, 1 indicates active
[3] "l4id" – the level 4 id (from database) for the selected sample
[4] "logc_max" – log base 10 of the maximum concentration tested
[5] "logc_min" – log base 10 of the minimum concentration tested
[6] "max_mean" – the maximum of the means at each concentration
[7] "max_med" – the maximum of the medians at each concentration
[8] "modl_ac10" – the activity concentration at 10% of the modeled top value (AC10)
[9] "modl_acb" – the activity concentration at baseline
[10] "modl_acc" – the activity concentration at cutoff
[11] "modl_ga" – the gain AC50
[12] "modl_gw" – the gain hill coefficient
[13] "modl_la" – the loss AC50
[14] "modl_lw" – the loss hill coefficient
[15] "modl" – the winning model
[16] "modl_rmse" – the root mean square error (RMSE)
[17] "modl_tp" – the modeled top of the curve
[18] "spid" – the sample id for the selected sample
[19] "tested" – whether the chemical was tested, 1 indicates tested
[20] "zscore" – the zscore of AC50 values based on the chemical-specific cytotoxicity distribution file

NOTE: For all matrix files, concentrations are given in log base 10 micromolar units.

All parameters beginning with "modl" are derived from the winning model. The complete set of parameters for all models is available in the MySQL database. NA values in the matrix files have different meanings, depending on the file. NA either means we did not test the chemical, or we could not compute the parameter. For example, when the constant model wins, we cannot compute a gain AC50. Similarly, if the Hill model wins, the loss AC50 is not applicable. NA in the hit-call matrix ("hitc") means the chemical did not get tested in the multiple concentration format. However, the chemical may have been tested in an initial screen at a single concentration and was not selected for further testing. The "tested" matrix indicates whether the chemical has been tested in an assay, and reflects both the single-concentration and multiple-concentration screening formats.

The hit-call matrix contains NA, 0, 1, and -1 values. "NA" indicates the chemical was not tested in the multiple-concentration screening format, "0" indicates the chemical was determined inactive, "1" indicates the chemical was determined active, and "-1" indicates that the activity could not be determined. Only chemicals with less than 4 concentrations of data received the "-1" designation.

The parameters for the winning model are given regardless of hit-calling; therefore, many inactive chemicals have a gain AC50 value in the "modl_ga" file.

The other two data files contain the cautionary flags for all the selected samples in the matrix files and the cytotoxicity distribution by chemical. The flag file contains many database id values, basic chemical information, the assay endpoint name (aenm), and the flag information. The flag information includes the flag database id (l6_mthd_id), the flag output (flag), and the flag value/unit (fval/fval_unit) when applicable. Not all flags have an associated value.

The cytotoxicity distribution file contains basic chemical information, the median (med) and MAD (mad) of gain AC50 across the cytotoxicity assays (in log base 10 micromolar units), and the number of cytotoxicity assays with an active hit-call (nhit). The global MAD (global_mad) is defined as the median of all the MAD values, excluding NA values. The cytotoxicity assays are indicated by the "burst_assay" field in the assay summary file. When a chemical hits less than two cytotoxicity assays, the cytotoxicity point (cyto_pt) is defined as 3, otherwise the cytotoxicity point is the cytotoxicity median (med) for the chemical. "cyto_pt_um" and "lower_bnd_um" are the cytoxicity point and the cytotoxicity point minus three times the global MAD in micromolar units, respectively.

The chemical summary file contains the mapping from "code" to CASRN (casn), DSSTox_GSID (chid), and chemical name (chnm) for the 1860 unique chemicals. The assay summary file contains a subset of the assay annotations to describe the 821 included assays. The complete set of assay annotation available through the MySQL database.

Detailed information about all of the parameters is available in the tcpl R package vignette, "Pipeline_Overview.pdf" (ToxCast Data Pipeline Overview).

For questions or concerns, please contact Monica Linnenbrink at: linnenbrink.monica@epa.gov.

```
#######################################################################
## R Script to produce December 2014 ToxCast Data Release
#######################################################################

library(tcpl)
library(data.table)
library(parallel)

## Write the matrix files and cytotoxicity distribution file
vars <- c("modl_ga", "hitc", "modl_tp", "modl_la", "modl", "max_mean",
          "modl_acc", "modl_acb", "modl_ac10", "max_med", "logc_max",
          "logc_min", "spid", "l4id", "modl_gw", "modl_lw", "fitc",
          "modl_rmse", "tested", "zscore")
res <-  mclapply(vars,
                 tcplVarMat,
                 clib = c("ph1v2_ph2_all_toxcast", "e1k_toxcast"),
                 srgx = "(?!^Tox21_1.*)(?!^Tox21_2.*)(?=^.*)",
                 odir = getwd(),
                 mc.preschedule = FALSE,
                 mc.cores = detectCores() - 1)

## Write the assay and chemical summary files
post <- format(Sys.Date(), "_%y%m%d.csv")
write.csv(tcplLoadAeidInfo("export_ready", 1),
          paste0("Assay_Summary", post),
          row.names = FALSE)
chid <- tcplLoadClib("clib",
                     val = c("e1k_880_toxcast",
                             "ph1v2_ph2_all_toxcast",
                             "ph1v1_toxcast"))
chem <- tcplLoadChem(field = "chid", val = unique(chid$chid))
chem <- chem[ ,
             unique(.SD),
             .SDcols = c("chid", "casn", "chnm"),
             key = "code"]
write.csv(chem, paste0("Chemical_Summary", post), row.names = FALSE)

## Write the flag file
names(res) <- vars
l4ids <- res[["l4id"]][!is.na(res[["l4id"]])]
write.csv(tcplPrepOtpt(tcplLoadData(lvl = 6L,
                                    fld = "l4id",
                                    val = l4ids)),
          paste0("AllResults_flags", post),
          row.names = FALSE)

#######################################################################
## End R script
#######################################################################
```